

Addis Ababa University

Master’s Thesis Proposal

**Name of Student:** Binalfew Kassa Mekonnen

**College:** Natural Sciences

**Department:** Computer Science

**Major:** Data and Web Engineering

**Name of Advisor:** Fekade Getrahun (PhD)

**Title of Thesis:** Knowledge Discovery from Free Text: Extraction of Violent Events

Table of Contents

[1. Introduction 2](#_Toc211761864)

[2. Motivation 3](#_Toc211761865)

[3. Statement of the Problem 5](#_Toc211761866)

[Problem 1: Lack of Structured Event Classification Frameworks 6](#_Toc211761867)

[Problem 2: Absence of Integrated Knowledge Base Architecture 6](#_Toc211761868)

[Problem 3: Inadequate Information Access Mechanisms 7](#_Toc211761869)

[Research Questions 8](#_Toc211761870)

[4. Objectives 9](#_Toc211761871)

[General Objective 9](#_Toc211761872)

[Specific Objectives 9](#_Toc211761873)

[Objective 1: Master Event Extraction Theory and Practice 9](#_Toc211761874)

[Objective 2: Develop Domain Expertise in Violence and Conflict Analysis 9](#_Toc211761875)

[Objective 3: Design a Hierarchical Violent Event Taxonomy 9](#_Toc211761876)

[Objective 4: Master Open-Source Technologies for Event Extraction and Knowledge Engineering 10](#_Toc211761877)

[Objective 5: Collect and Prepare Comprehensive Training Data 10](#_Toc211761878)

[Objective 6: Implement Event Extraction with Hierarchical Classification 11](#_Toc211761879)

[Objective 7: Design and Implement the Violent Event Knowledge Base 11](#_Toc211761880)

[Objective 8: Develop Natural Language Question-Answering Capability 11](#_Toc211761881)

[Objective 9: Develop and Validate the Integrated Prototype System 12](#_Toc211761882)

[Objective 10: Demonstrate Operational Utility and Impact 12](#_Toc211761883)

[5. Methodology 13](#_Toc211761884)

[Phase 1: Violent Event Taxonomy Development 13](#_Toc211761885)

[1.1 Literature Review and Framework Analysis 13](#_Toc211761886)

[1.2 Stakeholder Consultation and Requirements Analysis 13](#_Toc211761887)

[1.3 Taxonomic Dimensions Identification 14](#_Toc211761888)

[1.4 Hierarchical Structure Design 14](#_Toc211761889)

[1.5 Classification Criteria Specification 14](#_Toc211761890)

[1.6 Taxonomy Validation 14](#_Toc211761891)

[Phase 2: Data Collection and Annotation 15](#_Toc211761892)

[2.1 Data Source Identification and Collection 15](#_Toc211761893)

[2.2 Annotation Schema Development 15](#_Toc211761894)

[2.3 Annotation Process 15](#_Toc211761895)

[2.4 Data Preprocessing 16](#_Toc211761896)

[2.5 Dataset Partitioning 16](#_Toc211761897)

[Phase 3: Natural Language Processing Pipeline Implementation 16](#_Toc211761898)

[3.1 Core NLP Components 16](#_Toc211761899)

[3.2 Domain-Specific Enhancements 17](#_Toc211761900)

[3.3 Linguistic Feature Extraction 17](#_Toc211761901)

[Phase 4: Event Extraction and Attribute Identification 17](#_Toc211761902)

[4.1 Event Trigger Detection 17](#_Toc211761903)

[4.2 Argument Role Identification (5W Extraction) 17](#_Toc211761904)

[4.3 Multi-Event Handling 18](#_Toc211761905)

[4.4 Confidence Scoring 18](#_Toc211761906)

[Phase 5: Hierarchical Event Classification 18](#_Toc211761907)

[5.1 Feature Engineering for Classification 18](#_Toc211761908)

[5.2 Hierarchical Classification Strategy 19](#_Toc211761909)

[5.3 Classification Algorithm Selection 19](#_Toc211761910)

[5.4 Handling Classification Challenges 19](#_Toc211761911)

[5.5 Classification Evaluation 20](#_Toc211761912)

[Phase 6: Knowledge Base Design and Implementation 20](#_Toc211761913)

[6.1 Data Model Design 20](#_Toc211761914)

[6.2 Storage Architecture 20](#_Toc211761915)

[6.3 Indexing Strategies 21](#_Toc211761916)

[6.4 Data Integration Pipeline 21](#_Toc211761917)

[6.5 Knowledge Base Schema 21](#_Toc211761918)

[6.6 Scalability and Performance 21](#_Toc211761919)

[Phase 7: Question-Answering System Development 22](#_Toc211761920)

[7.1 Question Understanding Module 22](#_Toc211761921)

[7.2 Query Generation Module 22](#_Toc211761922)

[7.3 Query Execution and Answer Retrieval 22](#_Toc211761923)

[7.4 Answer Synthesis Module 22](#_Toc211761924)

[7.5 Dialogue Management 23](#_Toc211761925)

[7.6 User Interface Design 23](#_Toc211761926)

[7.7 Evaluation of Q&A System 23](#_Toc211761927)

[Phase 8: System Integration and Prototype Development 23](#_Toc211761928)

[8.1 Architecture Design 23](#_Toc211761929)

[8.2 Workflow Orchestration 24](#_Toc211761930)

[8.3 User Roles and Access Control 24](#_Toc211761931)

[8.4 Monitoring and Logging 24](#_Toc211761932)

[8.5 Testing Strategy 24](#_Toc211761933)

[Phase 9: Evaluation and Validation 24](#_Toc211761934)

[9.1 Extraction Evaluation 24](#_Toc211761935)

[9.2 Classification Evaluation 24](#_Toc211761936)

[9.3 Knowledge Base Evaluation 25](#_Toc211761937)

[9.4 Question-Answering Evaluation 25](#_Toc211761938)

[9.5 End-to-End Evaluation 25](#_Toc211761939)

[9.6 Error Analysis and Iteration 25](#_Toc211761940)

[Methodology Summary 25](#_Toc211761941)

[6. Related Work 26](#_Toc211761942)

[General Event Extraction Approaches 26](#_Toc211761943)

[Event Extraction from Large-Scale Knowledge Resources 26](#_Toc211761944)

[Event Extraction from Social Media 27](#_Toc211761945)

[Violence-Specific Event Extraction Systems 27](#_Toc211761946)

[Chinese Event Extraction and Ontology Population 28](#_Toc211761947)

[Event Classification and Taxonomy Design 28](#_Toc211761948)

[Knowledge Representation and Question-Answering 29](#_Toc211761949)

[Research Positioning 30](#_Toc211761950)

[7. Scope and Limitations 31](#_Toc211761951)

[Scope 31](#_Toc211761952)

[In Scope: 31](#_Toc211761953)

[Out of Scope: 32](#_Toc211761954)

[Limitations 33](#_Toc211761955)

[Data Limitations: 33](#_Toc211761956)

[Technical Limitations: 33](#_Toc211761957)

[Linguistic Limitations: 33](#_Toc211761958)

[Taxonomic Limitations: 34](#_Toc211761959)

[Knowledge Base Limitations: 34](#_Toc211761960)

[Question-Answering Limitations: 34](#_Toc211761961)

[Operational Limitations: 35](#_Toc211761962)

[Ethical Limitations: 35](#_Toc211761963)

[8. Application of Results 36](#_Toc211761964)

[Operational Early Warning and Crisis Response 36](#_Toc211761965)

[African Union Continental Early Warning System (AU-CEWS) 36](#_Toc211761966)

[Regional Economic Communities (RECs) Early Warning Systems 37](#_Toc211761967)

[National Early Warning Centers 37](#_Toc211761968)

[Humanitarian Response and Protection 37](#_Toc211761969)

[Humanitarian Organizations 37](#_Toc211761970)

[International Committee of the Red Cross (ICRC) 37](#_Toc211761971)

[Security and Defense Analysis 38](#_Toc211761972)

[African Union Peace Support Operations 38](#_Toc211761973)

[National Defense and Intelligence Services 38](#_Toc211761974)

[International Partners 38](#_Toc211761975)

[Conflict Research and Analysis 38](#_Toc211761976)

[Academic Research Community 38](#_Toc211761977)

[Think Tanks and Policy Institutes 39](#_Toc211761978)

[Conflict Early Warning and Forecasting Research 39](#_Toc211761979)

[Governance and Rule of Law 39](#_Toc211761980)

[Transitional Justice Mechanisms 39](#_Toc211761981)

[International Criminal Court (ICC) and Hybrid Courts 39](#_Toc211761982)

[National Justice Systems 39](#_Toc211761983)

[Development and Peacebuilding 39](#_Toc211761984)

[Development Organizations 39](#_Toc211761985)

[Peacebuilding Organizations 40](#_Toc211761986)

[Media and Public Information 40](#_Toc211761987)

[Journalists and Media Organizations 40](#_Toc211761988)

[Fact-Checking Organizations 40](#_Toc211761989)

[Public Communication and Civic Education 40](#_Toc211761990)

[Advanced Analytical Applications 40](#_Toc211761991)

[Network Analysis 40](#_Toc211761992)

[Temporal Pattern Mining 41](#_Toc211761993)

[Predictive Analytics Foundation 41](#_Toc211761994)

[System Scalability and Adaptation 41](#_Toc211761995)

[Adaptation to Other Event Types 41](#_Toc211761996)

[Geographic Extension 41](#_Toc211761997)

[Temporal Extension 41](#_Toc211761998)

[Operational Impact Assessment 42](#_Toc211761999)

[Quantifiable Benefits 42](#_Toc211762000)

[Qualitative Benefits 42](#_Toc211762001)

[Long-Term Strategic Value 42](#_Toc211762002)

[Policy Evidence Base 42](#_Toc211762003)

[Academic Contribution 43](#_Toc211762004)

[Capacity Building 43](#_Toc211762005)

[Ethical Considerations and Responsible Use 43](#_Toc211762006)

[9. Annexes 44](#_Toc211762007)

[Project Timeline 44](#_Toc211762008)

[10. References 46](#_Toc211762009)

# 1. Introduction

In the current digital era, an unprecedented volume of information is disseminated through the Internet, with many of these materials distributed in unstructured free text formats via websites, blogs, news wires, social media platforms, and mobile applications. Natural language processing (NLP) tools, which transcend simple keyword-based retrieval, are increasingly being deployed for autonomous knowledge discovery from the vast repositories of textual data available online. This trend is accelerating as organizations recognize the strategic value of converting unstructured narratives into actionable intelligence [1].

The exponential growth of digital content sources presents both opportunities and challenges for decision-making processes. An omnipresent obstacle is the inherently unstructured nature of textual data - information encoded in human-understandable natural language that lacks explicit machine-interpretable semantics. This characteristic severely constrains the automation of critical information retrieval (IR) and information extraction (IE) processes, particularly when processing massive datasets [2]. Moreover, the challenge extends beyond mere extraction; the extracted information must be systematically organized, classified, and made queryable to deliver genuine value to end-users.

Text Mining (TM) addresses these challenges through sophisticated information learning from pre-processed free text, employing techniques such as part-of-speech identification, stemming, and semantic analysis [3]. Through text mining coupled with Natural Language Processing (NLP), information is extracted from diverse sources and transformed into structured representations suitable for database storage, querying, and analysis. A particularly valuable form of knowledge extractable through TM is the **event** - a complex construct that can be represented as a structured combination of attributes (actors, actions, locations, temporal markers, and circumstances) derived from empirical observations within texts [4].

**Event extraction from unstructured text offers multifaceted benefits for information systems**. Beyond simple extraction, however, lies the critical challenge of **event classification and organization**. Events do not exist in isolation; they exhibit hierarchical relationships, taxonomic structures, and semantic connections that must be captured to enable sophisticated analysis. For instance, an “armed assault” is a specific type of “violent attack,” which itself belongs to the broader category of “armed violence.” Understanding these hierarchical relationships is essential for aggregating statistics, identifying patterns, and enabling users to query events at varying levels of granularity. Additionally, organizing extracted events into a **knowledge base** - a structured repository with semantic relationships and querying capabilities - transforms raw extracted data into a powerful analytical resource [5].

The extraction and classification of events combine knowledge from multiple disciplines including linguistics, computer science, data mining, artificial intelligence, and ontology engineering. Event extraction is commonly viewed as the text mining-aided identification and structuring of complex entity relationships, performed through a pipeline of natural language processing operations. It represents an advanced form of information extraction, typically domain-specific and resulting in detailed, structured outputs that can populate knowledge bases and support sophisticated query mechanisms [5].

**The development of event taxonomies** - hierarchical classification frameworks that define relationships among event types - is crucial for systematic event organization. Taxonomies enable consistent categorization, facilitate cross-dataset comparisons, and provide the semantic structure necessary for intelligent querying. When combined with knowledge base technologies, event taxonomies enable **question-answering systems** that allow users to retrieve specific information through natural language queries rather than requiring technical database expertise. This capability is particularly vital in operational contexts where analysts and decision-makers need rapid access to specific event information under time-critical conditions.

This research addresses the complete pipeline from raw text to queryable knowledge: extracting violent events from unstructured text, classifying them within a hierarchical taxonomy, organizing them into a structured knowledge base, and enabling intelligent question-answering capabilities. This end-to-end approach ensures that extracted information delivers maximum value to stakeholders involved in conflict analysis, early warning, and crisis response.

# 2. Motivation

At the Situation Monitoring Center of the African Union Continental Early Warning System (AU-CEWS), the operational mandate encompasses extensive data mining and intelligence analysis activities aimed at supporting timely, informed decision-making on conflict prevention and crisis response. The Africa Media Monitor tool successfully aggregates and clusters news articles from diverse sources; however, **the transformation of unstructured textual narratives into reliable, categorized, and queryable structured information remains a formidable challenge**. This difficulty stems from multiple factors: the massive volume of daily news production, the linguistic and structural heterogeneity of sources (including multiple languages, varied journalistic styles, and inconsistent reporting formats), and the complexity of violent events themselves, which often involve multiple actors, locations, and causal factors.

News articles arrive from varied channels - mainstream media outlets, local reporters, wire services, and increasingly from social media sources - each with different levels of detail, accuracy, and contextual information. Field reporters contribute valuable ground-level observations but face constraints in systematically documenting all incident attributes using standardized templates. Consequently, the majority of violent event information remains locked in narrative form, accessible only through manual reading and analysis - a process that cannot scale to meet the real-time demands of continental early warning systems.

**The critical gap lies not merely in extraction but in systematic organization and accessibility of event information**. Even when events are successfully extracted, they must be classified into meaningful categories to enable pattern recognition, trend analysis, and comparative assessments. For instance, distinguishing between different forms of political violence (coups, assassinations, election violence, repression of protests) versus criminal violence (gang conflicts, kidnappings, armed robberies) versus communal violence (ethnic clashes, land disputes, resource conflicts) is essential for tailoring appropriate responses. Furthermore, understanding hierarchical relationships among event types - recognizing that “suicide bombing” is a specific form of “terrorist attack,” which itself is a category within “political violence” - enables analysis at multiple levels of granularity.

The African Union is actively pursuing initiatives to automate the extraction and classification of violent events including homicides, injuries, kidnappings, arrests, hostage-taking, arson, armed assaults, riots, bombings, and various forms of political violence. The diversity of actors involved - state security forces, rebel groups, criminal organizations, ethnic militias, terrorist networks, and civilian mobs - coupled with the complexity of motivations and contexts, makes **systematic categorization** imperative for effective response. The AU’s mandate to assist victims, provide early warning, and support conflict resolution requires not just knowing that violence occurred, but understanding **what type of violence, perpetrated by whom, against whom, where, when, and under what circumstances**.

**The need for a hierarchical event taxonomy** emerges from operational requirements for event aggregation and reporting. Decision-makers require the flexibility to query events at different levels of specificity: “all violent events in the Sahel region” (broad), “terrorist attacks in West Africa” (intermediate), or “suicide bombings by JNIM in Mali” (specific). Without a structured taxonomy defining these relationships, such multi-level querying is impossible. Moreover, a well-designed taxonomy enables **statistical aggregation and trend analysis**: tracking whether overall violence is increasing or whether specific subcategories (e.g., election-related violence) are escalating while others decline.

**The imperative for a queryable knowledge base** stems from the operational reality that different users have different information needs and technical capabilities. Policy analysts need summary statistics and trends; humanitarian responders require detailed information about specific incidents affecting particular populations; early warning specialists search for patterns and precursors. A knowledge base architecture with **question-answering capabilities** democratizes access to extracted information, allowing users to formulate queries in natural language (“What kidnappings occurred in border regions last month?”) rather than requiring SQL expertise or manual database navigation.

Previous research by Taye Abdulkadir [8] made significant progress in extracting 5W characteristics (Who, What, Whom, Where, When) from news text using linguistically-informed methods and machine learning. However, several limitations constrain the operational deployment of that work: (1) limited training data resulted in suboptimal accuracy, (2) the use of Java rather than Python limited access to state-of-the-art machine learning libraries, (3) **no systematic event taxonomy was developed, preventing hierarchical classification and multi-level querying**, (4) **no knowledge base architecture was implemented, limiting the utility of extracted information**, and (5) **no user-facing query mechanism was provided, requiring technical expertise to access extracted data**.

This research addresses these gaps comprehensively by: (1) utilizing larger, more diverse training datasets, (2) employing Python-based machine learning frameworks with superior performance, (3) **developing a domain-specific hierarchical taxonomy of violent events contextualized for African conflicts**, (4) **designing and implementing a knowledge base architecture that stores extracted events with their taxonomic classifications and semantic relationships**, and (5) **creating an intelligent question-answering system that enables natural language querying of the knowledge base**. This integrated approach transforms event extraction from an academic exercise into an operational capability that directly supports AU-CEWS mandate.

The anticipated impact extends across multiple dimensions. **Operationally**, the system will reduce the time from event occurrence to actionable intelligence from days to hours or minutes. **Analytically**, the structured knowledge base will enable sophisticated pattern recognition, identification of escalation indicators, and evidence-based forecasting. **Strategically**, the accumulated historical knowledge base will support long-term conflict analysis, evaluation of intervention effectiveness, and development of context-specific prevention strategies. The question-answering capability will make this intelligence accessible to diverse stakeholders without requiring data science expertise, thereby democratizing access to critical information for decision-making.

While acknowledging limitations - the system targets English-language text, focuses on reported rather than predicted events, and requires ongoing maintenance of the taxonomy as new violence patterns emerge - this research represents a significant advancement in automated violent event intelligence for early warning and crisis response in the African context.

# 3. Statement of the Problem

Reports of violent incidents pervade contemporary news media, generating continuous streams of unstructured textual narratives about conflicts, attacks, riots, and other forms of violence. Manually processing these diverse sources to identify, extract, and analyze violent events is resource-intensive, time-consuming, and cannot achieve the speed and scale required for effective early warning and crisis response. While automated event extraction offers a solution, the challenge extends far beyond simple information extraction to encompass **systematic classification, structured organization, and intelligent retrieval** of event information.

The core problems addressed by this research can be articulated across three interconnected dimensions:

## Problem 1: Lack of Structured Event Classification Frameworks

Current approaches to violent event extraction often produce isolated event records without systematic categorization. This absence of structured classification creates multiple operational deficiencies:

* **Inconsistent categorization**: Without a standardized taxonomy, different analysts may classify the same event differently (e.g., is a politically-motivated killing classified as “assassination,” “targeted killing,” “political violence,” or “homicide”?), preventing consistent analysis and comparison.
* **Inability to aggregate at multiple granularities**: Decision-makers need flexibility to query events at varying levels of specificity. A regional commander may ask “what violent events occurred in my area of responsibility?” while a policy analyst needs “how many terrorist bombings occurred continent-wide?” Without hierarchical taxonomies defining relationships like “suicide bombing” ⊂ “bombing” ⊂ “explosive attack” ⊂ “terrorist attack” ⊂ “political violence,” such multi-level aggregation is impossible.
* **Loss of semantic relationships**: Events exist within taxonomic structures where understanding relationships enhances analytical insight. Knowing that “ethnic massacre” and “religious pogrom” are both forms of “communal violence” enables comparative analysis and pattern recognition that isolated event records cannot support.
* **Impediments to trend analysis**: Tracking whether specific forms of violence are escalating or declining requires consistent categorization over time. Ad-hoc classification schemes prevent longitudinal analysis and evidence-based forecasting.

## Problem 2: Absence of Integrated Knowledge Base Architecture

Even when events are successfully extracted with full 5W1H attributes, storing them in simple databases without semantic structure severely limits their analytical utility:

* **No relationship preservation**: Extracted events are stored as isolated records without capturing relationships such as “this attack was retaliation for that earlier incident” or “these three events were coordinated actions by the same group.” Knowledge bases with semantic relationship modeling can preserve these connections.
* **Inefficient querying**: Traditional database structures require users to know exact field names, table structures, and query syntax. This technical barrier prevents many potential users (policy analysts, humanitarian workers, military commanders) from accessing the information they need.
* **Limited cross-referencing**: Violent events often involve recurrent actors, locations, and patterns. Without a knowledge base architecture that maintains these cross-references, identifying that the same organization perpetrated attacks in multiple countries or that violence clusters along specific borders becomes computationally expensive and analytically cumbersome.
* **Lack of temporal reasoning**: Understanding event sequences, escalation patterns, and causal chains requires temporal reasoning capabilities beyond simple chronological sorting. Knowledge bases can encode temporal relationships (before/after, simultaneous, periodic) that enable sophisticated analysis.

## Problem 3: Inadequate Information Access Mechanisms

The ultimate value of extracted event information depends on users’ ability to retrieve relevant data efficiently:

* **Technical expertise barriers**: Requiring users to formulate SQL queries or navigate complex database schemas prevents non-technical stakeholders from accessing extracted information. An analyst should be able to ask “What kidnappings targeted aid workers in the Sahel region during the past quarter?” in natural language, not construct complex JOIN operations.
* **Inflexible query patterns**: Pre-built dashboards and reports provide fixed views of data but cannot accommodate the diverse, context-specific questions that arise in operational environments. Different crises demand different information, requiring adaptive query capabilities.
* **No natural language interface**: Human analysts think and communicate in natural language, not database query languages. The cognitive overhead of translating information needs into technical queries slows analysis and introduces errors.
* **Limited answer synthesis**: Users often need not just data retrieval but answer synthesis - aggregating information from multiple events, identifying patterns, and presenting coherent responses to analytical questions. Traditional database queries return raw records, requiring additional manual synthesis.

## Research Questions

To address these interconnected problems, this research investigates the following questions:

1. **Taxonomy Design**: What hierarchical taxonomy structure best captures the diversity of violent events in the African context while enabling multi-level categorization and querying? How should parent-child relationships be defined to balance specificity with generalizability? What taxonomic dimensions (actor type, attack method, target type, political context) should be encoded?
2. **Classification Methodology**: What hybrid approaches combining machine learning, linguistic analysis, and domain knowledge can most accurately assign extracted events to appropriate taxonomic categories? How can supervised learning be optimized when training data is limited? Can hierarchical classification methods that predict coarse categories before fine-grained subcategories improve accuracy?
3. **Knowledge Base Architecture**: What semantic data model best represents violent events with their 5W1H attributes, taxonomic classifications, and inter-event relationships? Should ontology languages like OWL be employed, or are graph databases more suitable? How should temporal and spatial dimensions be indexed for efficient querying?
4. **Question-Answering Implementation**: What natural language processing techniques enable accurate interpretation of user queries and mapping to knowledge base queries? How can the system handle ambiguous questions, resolve references, and synthesize multi-event answers? What interface design best serves users with varied technical backgrounds and operational contexts?
5. **Evaluation Metrics**: How should the integrated system be evaluated? Beyond extraction accuracy, how do we measure taxonomic classification accuracy, knowledge base query efficiency, question-answering correctness, and ultimately, operational utility for decision-makers?

By addressing these questions, this research aims to deliver not merely an event extraction system but a **comprehensive event intelligence platform** that extracts, classifies, organizes, and makes accessible violent event information for operational decision-making in African early warning and crisis response contexts.

# 4. Objectives

## General Objective

The overarching objective of this research is to develop an integrated, end-to-end system for violent event intelligence that extracts events from unstructured news text, classifies them within a hierarchical taxonomy, organizes them into a semantically-structured knowledge base, and enables natural language question-answering for operational decision support. This system will identify and extract the complete 5W1H attributes (Who committed the act, What type of event occurred, Whom was affected, Where it took place, When it happened, and How it unfolded) of violent events, classify each event within a domain-specific taxonomy, and provide stakeholders with intuitive access to this intelligence through natural language querying.

## Specific Objectives

### Objective 1: Master Event Extraction Theory and Practice

Acquire comprehensive theoretical and practical knowledge of information extraction (IE) with particular emphasis on event extraction methodologies. This encompasses: - Studying state-of-the-art event extraction architectures including pipeline-based, joint-inference, and end-to-end neural approaches - Understanding linguistic foundations including semantic role labeling, dependency parsing, and discourse analysis as they pertain to event understanding - Investigating hybrid methodologies that combine rule-based linguistic processing with machine learning - Analyzing domain adaptation techniques for transferring event extraction models to conflict and violence domains - Reviewing evaluation frameworks and metrics specific to event extraction tasks

### Objective 2: Develop Domain Expertise in Violence and Conflict Analysis

Conduct extensive research into violent events and conflict dynamics, particularly in African contexts, to inform taxonomy development and feature engineering. This includes: - Systematic review of violence and conflict classification frameworks from ACLED (Armed Conflict Location & Event Data Project), UCDP (Uppsala Conflict Data Program), GDELT (Global Database of Events, Language, and Tone), and similar initiatives - Consultation with AU-CEWS domain experts, conflict analysts, and field reporters to understand operational classification needs and challenges - Analysis of African conflict patterns, actor typologies, violence modalities, and contextual factors that should inform taxonomy structure - Investigation of how event intelligence supports early warning, conflict prevention, and crisis response decision-making - Understanding of the ethical implications of automated violence monitoring and potential biases in news reporting

### Objective 3: Design a Hierarchical Violent Event Taxonomy

Develop a comprehensive, hierarchical taxonomy of violent events optimized for the African context and AU-CEWS operational requirements. The taxonomy will: - **Define multi-level hierarchical structure**: Establish 3-5 levels of classification from broad categories (e.g., “Political Violence,” “Criminal Violence,” “Communal Violence”) through intermediate types (e.g., “State Repression,” “Rebellion,” “Terrorism”) to specific event types (e.g., “Extrajudicial Killing,” “Ambush,” “Suicide Bombing”) - **Specify taxonomic dimensions**: Incorporate multiple classification axes including actor type (state/non-state), violence type (physical/structural), target type (combatant/civilian), scale (individual/mass), and political context (election-related, resource-driven, ethnic) - **Establish classification criteria**: Define clear, operationalizable criteria for assigning events to each category, with decision rules for ambiguous cases - **Ensure operational utility**: Design the taxonomy to support actual AU-CEWS workflows including report generation, statistical aggregation, and trend analysis - **Validate with domain experts**: Iteratively refine the taxonomy through consultation with conflict analysts and testing against historical event datasets - **Enable extensibility**: Structure the taxonomy to accommodate new event types as violence patterns evolve, particularly emerging threats like cyber-attacks on critical infrastructure or drone-based attacks

### Objective 4: Master Open-Source Technologies for Event Extraction and Knowledge Engineering

Learn and achieve proficiency in state-of-the-art open-source tools and frameworks essential for building the integrated system: - **NLP Frameworks**: Stanford CoreNLP for linguistic analysis (parsing, NER, coreference); spaCy for efficient processing; NLTK for specialized linguistic tasks - **Machine Learning Platforms**: Scikit-learn for classical ML algorithms; TensorFlow or PyTorch for deep learning models if employed; Weka for baseline supervised classification - **Knowledge Representation**: RDF/OWL for semantic modeling; graph databases (Neo4j) for relationship management; SQL databases for structured attribute storage - **Question-Answering Technologies**: BERT-based models for semantic question understanding; entity linking systems; query generation frameworks - **Development Frameworks**: Python scientific stack (NumPy, Pandas); web frameworks for interface development; RESTful API design for system integration

### Objective 5: Collect and Prepare Comprehensive Training Data

Assemble, annotate, and preprocess substantial training datasets to support supervised machine learning for both event extraction and taxonomic classification: - **Data collection**: Aggregate diverse news articles from AU-CEWS Africa Media Monitor covering various African regions, time periods, and violence types to ensure representativeness - **Multi-level annotation**: Develop annotation guidelines and annotate training examples at multiple levels: - **Extraction level**: Mark entities (actors, victims, locations), event triggers, and 5W1H attributes - **Classification level**: Assign each annotated event to appropriate taxonomic categories at all hierarchy levels - **Relationship level**: Identify coreferent entities, temporal sequences, and causal relationships - **Quality assurance**: Employ multiple annotators with inter-annotator agreement measurement; adjudicate disagreements; ensure balanced representation of event types - **Data preprocessing**: Clean text, handle encoding issues, normalize temporal and spatial references, resolve abbreviations and acronyms - **Dataset partitioning**: Create training, validation, and test splits with stratification to ensure all taxonomic categories are adequately represented - **Data augmentation**: If needed, employ techniques like paraphrasing or synthetic example generation to address class imbalance for rare event types

### Objective 6: Implement Event Extraction with Hierarchical Classification

Design and develop the core extraction and classification system that processes raw text and produces classified event records: - **NLP pipeline implementation**: Build preprocessing pipeline (tokenization, sentence splitting, POS tagging, NER, parsing, coreference resolution) - **5W1H extraction**: Implement algorithms to identify event triggers and extract Who (actors), What (event type), Whom (victims), Where (locations), When (temporal references), and How (methods/circumstances) - **Candidate generation and ranking**: Develop methods to generate candidate entities for each role and rank them using linguistic features and machine learning - **Hierarchical classification**: Implement multi-level classification that first assigns events to broad categories, then progressively refines to more specific subcategories, leveraging the taxonomic structure to constrain and guide predictions - **Feature engineering**: Design features capturing linguistic patterns, semantic roles, contextual indicators, and domain-specific signals of violence - **Model training and optimization**: Train classification models on annotated data; perform hyperparameter tuning; implement cross-validation to assess generalization - **Error analysis**: Systematically analyze misclassifications to identify patterns and iteratively improve the system

### Objective 7: Design and Implement the Violent Event Knowledge Base

Develop a semantically-structured knowledge base architecture that stores extracted and classified events with queryable relationships: - **Data model design**: Specify conceptual schema representing events, actors, locations, temporal relations, and taxonomic classifications - **Relationship modeling**: Define semantic relationships including actor participation, victim impact, temporal precedence, spatial proximity, and causal linkage - **Storage architecture**: Implement hybrid storage combining relational databases for structured attributes, graph databases for relationships, and full-text indexes for content search - **Indexing strategies**: Create multi-dimensional indexes on temporal, spatial, actor, and taxonomic dimensions to enable efficient querying - **Data integration**: Develop pipelines to ingest extracted events, resolve entity references, deduplicate similar events, and maintain data consistency - **Provenance tracking**: Maintain links to source documents, extraction confidence scores, and modification histories for transparency and validation - **Scalability considerations**: Design the architecture to handle millions of events accumulated over years of continuous monitoring

### Objective 8: Develop Natural Language Question-Answering Capability

Create an intelligent interface that allows users to query the knowledge base using natural language questions: - **Question understanding**: Implement NLP components that parse user questions, identify question type (Who/What/Where/When/How many), extract constraints (temporal, spatial, categorical), and determine required information - **Query generation**: Develop algorithms that translate natural language questions into formal database queries (SQL, SPARQL, or graph queries) against the knowledge base - **Answer synthesis**: Implement logic to aggregate results from multiple events, format responses appropriately (single answer, list, summary statistics, timeline), and present supporting evidence - **Interactive clarification**: Handle ambiguous questions by requesting user clarification (e.g., “Did you mean Sudan or South Sudan?”) - **Query optimization**: Employ caching, query planning, and indexing to ensure response times suitable for interactive use (<2-3 seconds) - **Natural language generation**: Produce fluent, grammatical answer texts rather than raw database records - **Evaluation framework**: Test the Q&A system with benchmark questions covering diverse information needs and measure accuracy, completeness, and user satisfaction

### Objective 9: Develop and Validate the Integrated Prototype System

Construct a functional prototype that demonstrates the complete pipeline from text input to question-answering: - **System integration**: Connect all components (NLP processing, extraction, classification, knowledge base, Q&A interface) into a coherent workflow - **User interface development**: Create interfaces for uploading news articles, monitoring extraction progress, browsing the knowledge base, and submitting natural language queries - **API development**: Implement programmatic interfaces allowing external systems to submit documents and retrieve event information - **Performance optimization**: Profile system performance and optimize bottlenecks to achieve acceptable processing speeds (target: <1 minute per article) - **Evaluation**: Conduct comprehensive evaluation including: - **Extraction accuracy**: Precision, recall, and F1-score for 5W1H attribute extraction - **Classification accuracy**: Hierarchical precision and recall at each taxonomy level - **Knowledge base quality**: Completeness, consistency, and relationship accuracy - **Q&A correctness**: Answer accuracy, completeness, and relevance - **Usability testing**: User studies with AU-CEWS analysts assessing practical utility - **Documentation**: Create technical documentation, user manuals, and deployment guides

### Objective 10: Demonstrate Operational Utility and Impact

Validate the practical value of the system for AU-CEWS operations and similar organizations: - **Case study development**: Apply the system to analyze recent crisis periods (e.g., specific conflicts or election cycles) and demonstrate how the system would have supported decision-making - **Comparative analysis**: Benchmark system outputs against manual analysis by domain experts to assess accuracy and time savings - **Stakeholder engagement**: Present results to AU-CEWS personnel and gather feedback on operational utility and needed refinements - **Deployment pathway**: Define requirements and processes for transitioning the prototype to operational deployment - **Future research identification**: Document limitations encountered and identify directions for future enhancement

By systematically achieving these objectives, this research will deliver a comprehensive violent event intelligence platform that advances both the academic state-of-the-art in event extraction and provides tangible operational capabilities for conflict analysis and early warning in the African context.

# 5. Methodology

This research employs a **design science research methodology** combined with systematic experimentation to develop and evaluate the violent event intelligence system. The approach integrates multiple methodological frameworks: natural language processing for text analysis, machine learning for pattern recognition and classification, knowledge engineering for semantic structuring, and human-computer interaction principles for question-answering interface design [7]. The methodology encompasses six interconnected phases executed iteratively with continuous refinement based on evaluation results.

## Phase 1: Violent Event Taxonomy Development

The foundation of systematic event classification is a well-designed, hierarchical taxonomy that captures the diversity of violent events while enabling consistent categorization and multi-level querying.

### 1.1 Literature Review and Framework Analysis

Conduct comprehensive review of existing violence classification frameworks including: - **ACLED (Armed Conflict Location & Event Data Project)**: Analyze ACLED’s taxonomy of political violence including battles, violence against civilians, riots, protests, and explosions/remote violence [16] - **UCDP (Uppsala Conflict Data Program)**: Study UCDP’s classification of organized violence, one-sided violence, and non-state conflict [17] - **GDELT (Global Database of Events, Language, and Tone)**: Examine GDELT’s CAMEO event coding scheme and its applicability to violence classification [18] - **Academic typologies**: Review violence taxonomies from political science and conflict studies literature - **Legal frameworks**: Consider international humanitarian law and international criminal law classifications of violence

### 1.2 Stakeholder Consultation and Requirements Analysis

Engage with AU-CEWS personnel to understand operational classification needs: - **Interview domain experts**: Conduct semi-structured interviews with conflict analysts, early warning specialists, and field coordinators to identify how they currently categorize and analyze violence - **Analyze reporting templates**: Examine existing incident reporting templates to understand what attributes and categories are currently captured - **Survey end-users**: Assess what types of queries and aggregations decision-makers most frequently need - **Review historical data**: Analyze patterns in past reported events to ensure the taxonomy covers actual event diversity

### 1.3 Taxonomic Dimensions Identification

Define multiple classification dimensions that jointly characterize violent events: - **Actor-based dimension**: State forces vs. rebel groups vs. militias vs. criminal organizations vs. terrorist networks vs. communal groups - **Violence-type dimension**: Direct physical violence vs. structural violence vs. threat of violence - **Target-based dimension**: Combatants vs. civilians vs. infrastructure vs. symbolic targets - **Method-based dimension**: Firearms vs. explosives vs. edged weapons vs. vehicles vs. chemical/biological vs. fire - **Scale dimension**: Individual incidents vs. mass violence vs. systematic campaigns - **Political-context dimension**: Election-related vs. resource-driven vs. ethnic/religious vs. criminal vs. state repression

### 1.4 Hierarchical Structure Design

Develop multi-level taxonomy with clear parent-child relationships:

**Level 1 (Broad Categories)**: - Political Violence - Criminal Violence  
- Communal Violence - State Violence Against Civilians

**Level 2 (Intermediate Types)** - Example under Political Violence: - Rebellion/Insurgency - Terrorism - Coups and Regime Change - Election Violence - Repression of Opposition

**Level 3-4 (Specific Event Types)** - Example under Terrorism: - Bombing (→ Suicide Bombing, Car Bombing, IED Attack) - Armed Assault (→ Coordinated Attack, Lone Actor Attack) - Hostage-Taking - Assassination

### 1.5 Classification Criteria Specification

For each taxonomic category, define: - **Necessary conditions**: What attributes an event must possess to belong to this category - **Sufficient conditions**: What attributes definitively place an event in this category - **Distinguishing features**: How to differentiate this category from similar categories - **Typical examples**: Prototypical instances of this event type - **Edge cases**: Handling of ambiguous events that span multiple categories

### 1.6 Taxonomy Validation

* **Expert validation**: Present taxonomy to AU-CEWS analysts and conflict experts for review
* **Historical event coding**: Apply taxonomy to classify 200-300 historical events and assess coverage and clarity
* **Inter-coder reliability**: Have multiple annotators independently classify the same events and measure agreement (Cohen’s Kappa > 0.75)
* **Iterative refinement**: Modify taxonomy based on validation results until acceptable coverage and agreement are achieved

## Phase 2: Data Collection and Annotation

High-quality, comprehensively annotated training data is essential for supervised machine learning.

### 2.1 Data Source Identification and Collection

* **Primary source**: News articles from AU-CEWS Africa Media Monitor covering diverse African regions and time periods
* **Supplementary sources**: Reputable international news agencies (Reuters, AFP, BBC, Al Jazeera), regional news outlets, and conflict databases
* **Temporal coverage**: Collect articles spanning 2-3 years to capture diverse event types and contexts
* **Geographic coverage**: Ensure representation from all major African regions (North Africa, West Africa, East Africa, Central Africa, Southern Africa)
* **Event diversity**: Target 1,500-2,000 news articles describing approximately 2,500-3,500 violent events across all taxonomic categories

### 2.2 Annotation Schema Development

Create detailed annotation guidelines specifying:

**Entity-level annotation**: - **Actors**: Organizations, individuals, or groups perpetrating violence (mark entity spans, resolve coreferences, link to actor databases if available) - **Victims**: Individuals, groups, or entities suffering from violence (with similar span and linking) - **Locations**: Geographic references from countries to specific villages (with geocoding where possible) - **Temporal expressions**: Dates, date ranges, and relative temporal references - **Weapons/methods**: Instruments and tactics used in violence

**Event-level annotation**: - **Event triggers**: Words or phrases explicitly indicating violent events (verbs like “killed,” “attacked,” “bombed”) - **5W1H attributes**: Structured extraction of Who (actor), What (event type), Whom (victim), Where (location), When (time), How (method) - **Event boundaries**: Determining whether text describes single event or multiple distinct events - **Taxonomic classification**: Assigning each event to appropriate categories at all taxonomy levels - **Relationships**: Marking temporal sequences (Event A preceded Event B), causal links, coordinated attacks, retaliation patterns

### 2.3 Annotation Process

* **Annotator training**: Train 3-4 annotators on the annotation schema through practice sessions and discussion of ambiguous cases
* **Pilot annotation**: Annotate 50-100 articles as a pilot, measure inter-annotator agreement, refine guidelines based on confusion patterns
* **Distributed annotation**: Divide remaining articles among annotators
* **Double annotation**: Have 20% of articles independently annotated by two annotators to continuously monitor agreement
* **Adjudication**: Senior annotator or subject matter expert resolves disagreements
* **Quality control**: Regular annotation review sessions and feedback loops

### 2.4 Data Preprocessing

* **Text cleaning**: Remove HTML artifacts, advertisements, navigation elements
* **Encoding normalization**: Convert all text to UTF-8, handle special characters
* **Sentence segmentation**: Split articles into sentences (Stanford CoreNLP sentence splitter)
* **Tokenization**: Break sentences into tokens (words, punctuation)
* **Date normalization**: Convert relative dates (“yesterday,” “last week”) to absolute dates based on article timestamp
* **Location normalization**: Map location mentions to standard geographic coordinates and hierarchical place names
* **Acronym expansion**: Maintain dictionary of common acronyms (e.g., “ECOWAS,” “AQIM”) and expand on first mention

### 2.5 Dataset Partitioning

* **Training set**: 70% of annotated data for model training
* **Validation set**: 15% for hyperparameter tuning and model selection
* **Test set**: 15% held out until final evaluation, ensuring no overlap with training/validation
* **Stratification**: Ensure all taxonomic categories represented in all splits proportionally

## Phase 3: Natural Language Processing Pipeline Implementation

Linguistic preprocessing transforms raw text into structured representations suitable for event extraction.

### 3.1 Core NLP Components

**Stanford CoreNLP Integration**: - **Tokenization and sentence splitting**: Segment text into analyzable units - **Part-of-speech tagging**: Label each word with grammatical category (noun, verb, adjective, etc.) - **Lemmatization**: Reduce words to base forms (e.g., “killed,” “killing,” “kills” → “kill”) - **Named Entity Recognition (NER)**: Identify and classify named entities (PERSON, ORGANIZATION, LOCATION, DATE) - **Constituency parsing**: Generate phrase-structure parse trees capturing grammatical relationships - **Dependency parsing**: Extract grammatical dependencies between words (subject, object, modifier relationships) - **Coreference resolution**: Identify when different noun phrases refer to the same entity (e.g., “the rebels” and “they” referring to the same group)

### 3.2 Domain-Specific Enhancements

* **Violence lexicon**: Compile comprehensive list of violence-related terms (verbs like “attacked,” “killed,” “kidnapped”; nouns like “bombing,” “massacre,” “ambush”)
* **Actor name recognition**: Enhance NER with African organization names, armed groups, ethnic groups not well-covered by standard models
* **Location recognition**: Improve recognition of African place names, including alternative spellings and local language names
* **Temporal expression normalization**: Resolve relative and context-dependent temporal references to absolute dates

### 3.3 Linguistic Feature Extraction

Extract features from parsed text to support event extraction and classification: - **Semantic roles**: Identify agent, patient, instrument, location, and time for violence-related predicates - **Syntactic patterns**: Detect common grammatical structures for violence reporting (e.g., “X killed Y in Z” pattern) - **Contextual embeddings**: Generate word and sentence embeddings using pre-trained models (e.g., BERT) to capture semantic context - **Event trigger indicators**: Identify words that signal violent events and their surrounding context

## Phase 4: Event Extraction and Attribute Identification

This phase implements algorithms to extract complete 5W1H event representations from linguistically processed text.

### 4.1 Event Trigger Detection

* **Pattern-based approach**: Use violence lexicon and syntactic patterns to identify candidate event triggers
* **Machine learning classification**: Train classifier to distinguish actual violence-related triggers from general uses of potentially ambiguous words
* **Context analysis**: Examine surrounding words and grammatical structure to confirm violent context

### 4.2 Argument Role Identification (5W Extraction)

**Who (Actor) Extraction**: - **Candidate generation**: Extract all PERSON and ORGANIZATION entities in syntactic vicinity of event trigger - **Role classification**: Use supervised classifier with features including: - Grammatical role (subject vs. object of violence verb) - Semantic role (agent vs. patient) - Contextual words (e.g., “forces,” “gunmen,” “militants” suggesting actor role) - Entity type and properties - **Actor linking**: Link extracted actors to knowledge base of known groups and individuals where possible

**Whom (Victim) Extraction**: - Similar pipeline to actor extraction but looking for patient/object roles - Special handling for mass victims (e.g., “civilians,” “protesters”) versus named individuals - Extraction of victim attributes (casualties counts, demographics)

**Where (Location) Extraction**: - Identify location entities and prepositional phrases indicating location - Resolve location ambiguities (multiple locations mentioned: which is event location vs. broader context?) - Geocoding: convert location names to coordinates using gazetteers and geocoding APIs - Handle varying geographic granularities (from village to country level)

**When (Temporal) Extraction**: - Extract temporal expressions using Stanford CoreNLP’s SUTime - Normalize to absolute dates considering article publication date - Handle temporal uncertainties and ranges - Extract event duration if mentioned

**What (Event Type) Extraction**: - Primary extraction from event trigger word (e.g., “bombing” suggests explosive attack) - Refined based on contextual information and full 5W picture - Initial classification at general level; fine-grained classification in Phase 5

**How (Method/Circumstances) Extraction**: - Identify weapon/instrument mentions (firearms, explosives, vehicles) - Extract tactical details (ambush, raid, coordinated attack) - Capture contextual circumstances (during protest, at checkpoint, in marketplace)

### 4.3 Multi-Event Handling

* **Event boundary detection**: Determine whether article describes single event or multiple distinct events
* **Event coreference**: Identify when different sentences describe the same event versus different related events
* **Cross-sentence information integration**: Aggregate information about same event scattered across multiple sentences

### 4.4 Confidence Scoring

Assign confidence scores to each extracted attribute based on: - Linguistic evidence strength (clear syntactic roles vs. ambiguous attachment) - Entity recognition confidence - Classification model probability scores - Number of supporting mentions

## Phase 5: Hierarchical Event Classification

Extracted events are classified into the taxonomic framework through multi-level supervised classification.

### 5.1 Feature Engineering for Classification

Design comprehensive feature sets capturing event characteristics:

**Lexical features**: - Event trigger word and its lemma - Presence of specific violence-related keywords - Weapon/method mentions - Actor and victim descriptors

**Semantic features**: - Word embeddings for event trigger and surrounding context - Sentence embeddings for event-describing sentences - Semantic similarity to prototypical examples of each category

**Structural features**: - Number of casualties (if mentioned) - Number of actors and victims - Geographic location and its characteristics - Temporal patterns (election period, holiday, historical anniversaries)

**Contextual features**: - Actor type (if recognized: state forces, rebel group, criminal gang) - Target type (civilian, military, infrastructure) - News source and framing - Surrounding events in the article (broader conflict context)

### 5.2 Hierarchical Classification Strategy

Rather than flat multi-class classification, employ hierarchical approach that mirrors taxonomy structure:

**Stage 1 - Top-level classification**: - Train classifier to distinguish broad categories: Political Violence vs. Criminal Violence vs. Communal Violence vs. State Violence - Uses general features and high-level patterns - Higher confidence in these coarse distinctions

**Stage 2 - Mid-level classification**: - For events classified as Political Violence, train second-stage classifier to distinguish: Rebellion vs. Terrorism vs. Coup vs. Election Violence vs. Repression - Conditional on top-level prediction, enabling focused feature sets - Similar cascaded classifiers for other top-level categories

**Stage 3 - Fine-grained classification**: - Within each mid-level category, train specific classifiers for detailed event types - Example: within Terrorism, classify as Bombing vs. Armed Assault vs. Kidnapping vs. Assassination - Then final layer: within Bombing, classify as Suicide Bombing vs. Car Bombing vs. IED

### 5.3 Classification Algorithm Selection

Experiment with multiple algorithms and select based on validation set performance: - **Random Forest**: Good baseline, handles diverse feature types, provides feature importance - **Support Vector Machines (SVM)**: Effective for high-dimensional feature spaces - **Gradient Boosting (XGBoost, LightGBM)**: Often achieves best performance on structured features - **Neural Networks**: Multi-layer perceptrons or more sophisticated architectures if data supports - **Ensemble methods**: Combine multiple classifiers through voting or stacking

### 5.4 Handling Classification Challenges

* **Class imbalance**: Employ SMOTE (Synthetic Minority Over-sampling), class weighting, or focal loss to handle rare event types
* **Ambiguous cases**: Allow multi-label classification when events genuinely span categories
* **Confidence calibration**: Calibrate probability outputs to reflect true classification confidence
* **Hierarchical consistency**: Ensure predictions respect hierarchy (can’t predict “Suicide Bombing” without also predicting “Bombing” and “Terrorism”)

### 5.5 Classification Evaluation

* **Accuracy metrics**: Precision, recall, F1-score at each taxonomic level
* **Hierarchical evaluation**: Hierarchical precision and recall measuring partial credit for ancestor categories
* **Confusion analysis**: Examine which categories are frequently confused to refine features or taxonomy
* **Error categorization**: Classify errors as extraction failures (wrong input to classifier) vs. true classification errors

## Phase 6: Knowledge Base Design and Implementation

Transform extracted and classified events into a semantically-structured, queryable knowledge base.

### 6.1 Data Model Design

**Conceptual schema**: - **Event entity**: Core object with attributes (unique ID, event type, timestamp, location coordinates, confidence scores, source document) - **Actor entity**: Organizations, groups, individuals involved as perpetrators - **Victim entity**: Individuals, groups, organizations suffering violence - **Location entity**: Places with hierarchical structure (village ⊂ district ⊂ province ⊂ country) - **Temporal entity**: Dates, date ranges, temporal relationships - **Taxonomic classification**: Linkages to taxonomy at all levels

**Relationships**: - **Participation**: Actor participates-in Event (with role: perpetrator, organizer, commander) - **Impact**: Event affects Victim (with impact type: killed, injured, displaced, kidnapped) - **Location**: Event occurs-at Location - **Temporal**: Event occurs-at Time; Event before/after Event - **Causality**: Event causes/triggers Event - **Coordination**: Event coordinated-with Event (simultaneous attacks) - **Retaliation**: Event retaliates-for Event - **Taxonomic**: Event instance-of EventType; EventType subclass-of EventType

### 6.2 Storage Architecture

Hybrid storage strategy combining strengths of different database paradigms:

**Relational database (PostgreSQL)**: - Stores structured event attributes (5W1H fields, confidence scores, metadata) - Enables efficient querying on specific attributes - Handles transactional consistency and ACID properties - PostGIS extension for spatial queries

**Graph database (Neo4j)**: - Models semantic relationships between events, actors, locations - Enables graph traversal queries (e.g., “all events connected by retaliation chains”) - Efficiently represents taxonomic hierarchy - Supports path-finding and network analysis

**Full-text search engine (Elasticsearch)**: - Indexes source document text for keyword search - Enables hybrid queries combining structured filters and text search - Provides relevant ranking and highlighting

### 6.3 Indexing Strategies

Create multi-dimensional indexes for efficient querying: - **Temporal index**: B-tree index on event dates; temporal range queries - **Spatial index**: R-tree or grid-based spatial index on location coordinates - **Taxonomic index**: Inverted index mapping each taxonomic category to member events - **Actor index**: Hash index on actor names with fuzzy matching support - **Composite indexes**: Combined indexes on frequent query patterns (e.g., location + time range + event type)

### 6.4 Data Integration Pipeline

Automated workflow to ingest extracted events into knowledge base: 1. **Entity resolution**: Identify when extracted entities refer to same real-world actors/locations across events 2. **Deduplication**: Detect when multiple articles describe the same event; merge redundant records 3. **Relationship inference**: Automatically infer some relationships (e.g., spatial proximity, temporal sequences) 4. **Quality validation**: Flag events with low confidence scores or missing critical attributes for manual review 5. **Provenance tracking**: Maintain links to source documents, extraction timestamps, system version 6. **Incremental update**: Support adding new events without full database rebuild

### 6.5 Knowledge Base Schema

Implement using semantic web standards where appropriate: - **RDF triples** for simple assertions (Event-123 has-location Nairobi) - **OWL ontology** defining taxonomy structure and relationship semantics - **SPARQL endpoint** enabling semantic queries for advanced users - **GraphQL API** providing flexible data access for application developers

### 6.6 Scalability and Performance

* **Partitioning**: Partition by date range to manage growth over years of continuous operation
* **Caching**: Cache frequent queries and aggregate statistics
* **Materialized views**: Pre-compute common aggregations (events per region per month)
* **Asynchronous processing**: Queue-based ingestion to handle burst loads during major crises

## Phase 7: Question-Answering System Development

Enable natural language access to the knowledge base through an intelligent Q&A interface.

### 7.1 Question Understanding Module

**Question parsing**: - Identify question type: Who/What/Where/When/How-many/List/Yes-No/Comparison - Extract question focus: What information is being requested? - Identify constraints: Temporal bounds, geographic scope, event categories, actors

**Named entity recognition in questions**: - Detect mentions of specific actors, locations, dates in user questions - Link entities to knowledge base identifiers

**Intent classification**: - Statistical query (count, sum, average) - Entity retrieval (specific events or actors) - Relationship query (connections between entities) - Temporal pattern (trends over time) - Spatial pattern (hotspots, clustering)

### 7.2 Query Generation Module

Translate natural language to formal database queries:

**Template-based approach**: - Maintain library of query templates for common question patterns - Fill template slots with entities and constraints from question - Example: “What [EVENT\_TYPE] occurred in [LOCATION] in [TIME]?” → SQL SELECT with WHERE clauses

**Semantic parsing approach**: - Use semantic grammar to map questions to logical forms - Logical forms translated to SPARQL or SQL depending on query complexity

**Neural seq2seq approach** (if resources permit): - Train encoder-decoder model on question-query pairs - Generates database queries end-to-end from questions - Requires substantial training data but handles novel questions better

### 7.3 Query Execution and Answer Retrieval

* **Query optimization**: Plan efficient execution considering indexes and joins
* **Timeout handling**: Set time limits for complex queries; return partial results if needed
* **Result ranking**: Order results by relevance, date, or confidence
* **Result limiting**: Return top-K results for large result sets with pagination

### 7.4 Answer Synthesis Module

Transform raw query results into natural language answers:

**Answer type adaptation**: - **Single entity**: “The attack occurred in Nairobi on March 15, 2024.” - **List**: “Three kidnapping events were reported: [list with details]” - **Count**: “47 violent events occurred in the Sahel region last month.” - **Aggregate statistics**: “Violence increased by 23% compared to the previous quarter.” - **Timeline**: Present events chronologically with brief descriptions - **Map visualization**: For spatial queries, show events on interactive map

**Natural language generation**: - Template-based generation for simple, predictable answers - Neural NLG models for complex answer synthesis if warranted - Include source attribution: “According to 3 news reports…”

**Supporting evidence**: - Link to source documents for verification - Display relevant text excerpts - Show confidence scores when appropriate

### 7.5 Dialogue Management

Handle multi-turn interactions: - **Context tracking**: Remember previous questions in conversation to resolve pronouns and implicit references - **Clarification requests**: Ask user to disambiguate when question is ambiguous (“Did you mean Sudan or South Sudan?”) - **Follow-up questions**: Support refinement queries (“Show me only suicide bombings from that list”) - **Exploratory dialogue**: Suggest related queries user might find interesting

### 7.6 User Interface Design

**Web-based interface**: - Text input for natural language questions - Autocomplete suggestions based on common queries - Result display with filtering, sorting, export options - Visualizations (charts, timelines, maps) for appropriate queries - Feedback mechanism for users to rate answer quality

### 7.7 Evaluation of Q&A System

* **Answer correctness**: Human evaluation of whether answers correctly address questions
* **Answer completeness**: Does answer provide all relevant information?
* **Query benchmark**: Create test set of 100+ diverse questions with gold-standard answers
* **Metrics**: Accuracy, F1-score (treating answer as set of facts), Mean Reciprocal Rank
* **User study**: AU-CEWS analysts use system for actual information needs; measure task completion, time savings, satisfaction

## Phase 8: System Integration and Prototype Development

Assemble all components into functional end-to-end system.

### 8.1 Architecture Design

* **Modular design**: Separate components with defined interfaces
* **API-first approach**: Each module exposes RESTful APIs
* **Microservices consideration**: Evaluate whether microservices architecture appropriate for scalability
* **Message queue**: Use RabbitMQ or Kafka for asynchronous processing

### 8.2 Workflow Orchestration

Define complete pipeline from article input to queryable knowledge: 1. Article ingestion (upload or automatic grabbing) 2. NLP preprocessing 3. Event extraction 4. Classification 5. Knowledge base integration 6. Available for querying

### 8.3 User Roles and Access Control

* **Administrator**: System configuration, user management
* **Data manager**: Upload articles, monitor extraction quality
* **Analyst**: Query knowledge base, generate reports
* **Viewer**: Read-only access to predefined reports

### 8.4 Monitoring and Logging

* **Processing logs**: Track each article through pipeline
* **Error logging**: Capture and categorize errors for debugging
* **Performance monitoring**: Track processing times, query latencies
* **Quality metrics dashboard**: Display extraction accuracy, classification performance, data coverage

### 8.5 Testing Strategy

* **Unit tests**: Test individual functions and methods
* **Integration tests**: Test component interactions
* **End-to-end tests**: Process sample articles through complete pipeline
* **Performance tests**: Measure scalability under load
* **User acceptance testing**: AU-CEWS personnel test with real use cases

## Phase 9: Evaluation and Validation

Comprehensive assessment of system performance and operational utility.

### 9.1 Extraction Evaluation

* Precision, recall, F1-score for each 5W1H attribute
* Entity-level evaluation and span-level evaluation
* Comparison to human annotations on test set

### 9.2 Classification Evaluation

* Hierarchical precision/recall at each taxonomy level
* Confusion matrix analysis
* Comparison to baseline classifiers
* Cross-validation to assess generalization

### 9.3 Knowledge Base Evaluation

* Coverage: percentage of events successfully integrated
* Consistency: detection of contradictions or duplications
* Relationship accuracy: correctness of inferred relationships
* Query performance: response time for representative queries

### 9.4 Question-Answering Evaluation

* Answer accuracy on benchmark question set
* User study with AU-CEWS analysts
* Comparison to manual information retrieval

### 9.5 End-to-End Evaluation

* Case study: analyze major crisis period and compare system output to expert analysis
* Time savings analysis: measure time for analysts to answer questions with vs. without system
* Decision impact assessment: qualitative evaluation of whether system improves decision quality

### 9.6 Error Analysis and Iteration

* Systematic analysis of errors at each stage
* Identification of improvement priorities
* Implementation of refinements based on findings
* Multiple evaluation cycles until acceptable performance achieved

## Methodology Summary

This comprehensive methodology integrates taxonomy design, machine learning, knowledge engineering, and natural language interfaces to deliver an end-to-end violent event intelligence system. The design science approach emphasizes iterative development with continuous evaluation and stakeholder feedback. The hybrid combination of linguistic analysis, supervised classification, and semantic knowledge representation addresses the complementary challenges of accurate extraction, systematic classification, and intelligent access to event information. Success requires careful attention to data quality, domain expertise integration, and alignment with operational requirements throughout the development process.

# 6. Related Work

Event extraction from diverse information sources has been the subject of extensive research spanning natural language processing, information extraction, knowledge representation, and database systems. This review examines relevant work across four key dimensions: general event extraction methodologies, violence-specific event extraction systems, event taxonomies and classification frameworks, and question-answering systems over structured knowledge.

## General Event Extraction Approaches

Event extraction methodologies can be broadly categorized into data-driven, knowledge-driven, and hybrid approaches [9]. **Data-driven methods** rely primarily on statistical machine learning trained on annotated corpora. These approaches learn patterns directly from data without requiring extensive manual rule engineering. **Knowledge-driven methods** utilize hand-crafted rules, linguistic patterns, and domain ontologies to identify and extract events. These methods benefit from linguistic and domain expertise but struggle with coverage and adaptability. **Hybrid approaches** combine the strengths of both paradigms, using linguistic analysis to constrain and guide machine learning, while employing statistical methods to achieve robustness and adaptability [2]. This research adopts a hybrid methodology, leveraging Stanford CoreNLP for linguistic preprocessing while employing supervised machine learning for classification tasks.

Hogenboom et al. [2,5] provide comprehensive surveys of event extraction methods from text for decision support systems. They identify key challenges including: defining what constitutes an event, handling the complexity of natural language, integrating information scattered across multiple sentences or documents, and representing extracted events in structured, queryable formats. Their work emphasizes that event extraction is not merely an NLP problem but requires careful consideration of knowledge representation, user requirements, and system usability for decision support applications.

## Event Extraction from Large-Scale Knowledge Resources

Several research efforts have focused on extracting events from structured or semi-structured knowledge sources. Suchanek et al. [10] developed YAGO, a large-scale event ontology constructed from Wikipedia and WordNet. Using rule-based and heuristic approaches, they automatically extracted event facts from approximately 1.6 million Wikipedia articles as of January 2007. YAGO treats each Wikipedia article as describing a single entity or event and employs infobox information and category structures to extract structured knowledge. While impressive in scale, YAGO’s focus on historical events and encyclopedic coverage differs from the real-time news monitoring requirements of early warning systems.

Hienert and Luciano [11] built upon similar principles to extract historical events spanning approximately 2,500 years from multilingual Wikipedia articles. They employed the DBpedia ontology type system for event identification and used semantic parsing to structure event information according to the LODE (Linking Open Descriptions of Events) event model. Their work demonstrates the value of standardized event models but highlights challenges in extracting from semi-structured sources that may not translate directly to free-text news articles.

## Event Extraction from Social Media

Social media platforms present both opportunities and challenges for event extraction. Becker et al. [13] developed query-oriented solutions for retrieving social media documents about planned events across multiple platforms. They exploited user-contributed structured data (event titles, times, locations posted on platforms like Eventful) combined with discussions and reactions on Twitter and Facebook. Their approach handles planned events rather than the unexpected violent incidents targeted in this research, but their cross-platform information integration techniques offer valuable insights.

Magnuson et al. [12] proposed an event recommendation system for Twitter users that identifies co-located Twitter activity associated with past events to drive geographic recommendations using item-based collaborative filtering. They relied on Eventbrite as their event source. These social media approaches highlight the potential of real-time monitoring but face challenges of noise, veracity, and biased geographic/demographic coverage that make them unsuitable as primary sources for formal early warning systems without substantial validation.

Aratefeh and Khreich [9] surveyed techniques specifically for event detection in Twitter, identifying challenges unique to social media including short text length, informal language, high noise levels, and the need for real-time processing. Their work categorizes approaches into supervised, unsupervised, and hybrid methods, and discusses the role of topic modeling, classification, and clustering. While this research targets traditional news media rather than social media, some techniques (particularly for handling informal or fragmented text) may prove applicable.

## Violence-Specific Event Extraction Systems

The extraction of violent events specifically has attracted research attention due to operational importance for security and conflict monitoring.

Piskorski et al. [14] developed NEXUS (News cluster Event eXtraction Utilizing language Structures), which extracts security-related events using the Ontology of Politically Motivated Violent Events (PMVE). NEXUS employs keyword-based heuristics to identify security-relevant news, then applies linguistic analysis and ontology mapping to extract structured information about violent events and associated entities (people, organizations). The PMVE ontology provides a domain-specific framework for representing violence, offering a valuable reference for taxonomic design in this research. However, NEXUS focuses on European contexts and may require adaptation for African conflict patterns.

Tanev et al. [1] developed real-time news event extraction capabilities for global crisis monitoring, addressing the challenge of processing large volumes of multilingual news for emerging crises. Their system combines pattern-based event recognition with geographical clustering to identify crisis locations and track evolving situations. While their focus on real-time processing differs from this research’s emphasis on systematic classification and knowledge base construction, their scalability techniques offer valuable lessons.

**The work most directly related to this research is that of Taye Abdulkadir and Sungkur [8]**, who developed a system for extracting violent events in the African context with focus on identifying 5W characteristics (Who, What, Whom, Where, When). Their hybrid approach combined linguistic analysis using Stanford CoreNLP with machine learning classification using Weka. They addressed challenges specific to African conflicts including diverse actor types, complex ethnic and political contexts, and reporting in multiple languages. **However, several limitations constrain the operational utility of their work: (1) relatively small training dataset limiting classification accuracy, (2) lack of a systematic hierarchical taxonomy preventing structured classification and multi-level querying, (3) absence of a knowledge base architecture limiting information integration and analysis, and (4) no user-facing query interface requiring technical database skills to access extracted information.** This research directly addresses these limitations through larger-scale data annotation, explicit taxonomy development, knowledge base implementation, and natural language question-answering capabilities.

## Chinese Event Extraction and Ontology Population

Wang and Zhao [15] proposed a framework for extracting 5W1H semantic elements from Chinese news to populate event ontologies. They developed the News Ontology Event Model (NOEM) representing semantic components and inter-event relationships. Extracted semantic elements are represented as RDF triples and automatically imported into NOEM as instances. Their work demonstrates the value of ontology-based event representation for enabling semantic reasoning and querying, though their focus on Chinese poses different linguistic challenges than English. Their ontology-population approach offers insights for knowledge base integration in this research.

## Event Classification and Taxonomy Design

While event extraction has received substantial research attention, the development of robust, operational event classification frameworks specifically for violence has been driven primarily by data collection organizations rather than academic research.

**The Armed Conflict Location & Event Data Project (ACLED)** [16] has developed one of the most comprehensive and widely-used violence taxonomies for Africa and beyond. ACLED classifies events into six primary types: Battles, Explosions/Remote Violence, Violence Against Civilians, Protests, Riots, and Strategic Developments. Each primary type has sub-event categories (e.g., Violence Against Civilians includes Attack, Abduction/Forced Disappearance, Sexual Violence). ACLED’s taxonomy has evolved through years of operational data collection and has been validated through extensive use by researchers and practitioners. **This research draws heavily on ACLED’s taxonomic structure while adapting and extending it for automated extraction and classification requirements.**

**The Uppsala Conflict Data Program (UCDP)** [17] provides another influential framework, distinguishing between State-Based Conflict, Non-State Conflict, and One-Sided Violence. UCDP’s definitions emphasize organized actors and fatality thresholds, making it valuable for conflict research but potentially missing lower-intensity violence relevant to early warning. Their careful definitional work regarding what constitutes organized violence versus other forms provides important theoretical grounding.

**The Global Database of Events, Language, and Tone (GDELT)** [18] employs the CAMEO (Conflict and Mediation Event Observations) event coding framework, a comprehensive taxonomy covering cooperative and conflictual events. While GDELT achieves impressive scale through automated processing, its very broad scope (all events, not just violence) and relatively coarse-grained automated extraction make it less suitable as a direct model for specialized violent event extraction. However, GDELT’s architecture for continuous global monitoring at scale offers valuable technical insights.

Academic research on violence typologies includes Nordås and Davenport’s work on different forms of state repression, the Political Terror Scale’s categorization of government violence, and various terrorism databases’ classification schemes. These sources inform taxonomic design but typically focus on specific violence subtypes rather than providing comprehensive frameworks.

**A critical gap in existing work is the integration of taxonomic classification into automated extraction systems.** Most violence taxonomies were designed for human coding and may not translate directly to automated classification features. This research addresses this gap by co-designing the taxonomy and automated classification system to ensure the taxonomy’s categories are distinguishable based on linguistic features extractable from news text.

## Knowledge Representation and Question-Answering

The utility of extracted events depends critically on how they are represented and made accessible to users.

**Semantic Web and ontology-based approaches** have been proposed for event representation. The Event Ontology (Raimond et al.) and LODE (Linking Open Descriptions of Events) provide standardized vocabularies for describing events in RDF/OWL, enabling semantic reasoning and cross-dataset linking. These frameworks emphasize interoperability and machine-readable semantics. This research evaluates whether full semantic web standards are warranted or whether lighter-weight graph database representations suffice for operational requirements.

**Question-answering over structured knowledge** has advanced significantly with neural approaches. Systems like IBM Watson pioneered sophisticated Q&A over heterogeneous knowledge sources. More recently, semantic parsing approaches (Berant et al.) map natural language to formal queries (SQL, SPARQL) through learned compositional semantics. Neural semantic parsing using sequence-to-sequence models (Dong and Lapata) can generate queries end-to-end from natural language questions.

For event-oriented Q&A specifically, Voskarides et al. developed question answering over temporal event knowledge graphs, addressing challenges of temporal reasoning and event relationships. Their work on interpreting temporal constraints in questions and generating appropriate queries offers relevant techniques for this research.

**Knowledge base question answering (KBQA)** research has produced various architectural patterns: pipeline approaches (question analysis → query generation → answer synthesis) versus end-to-end neural approaches. This research adopts a hybrid pipeline approach that balances interpretability and control (valuable for operational systems) with the pattern recognition capabilities of learned models.

## Research Positioning

This research builds upon and extends prior work in several key dimensions:

1. **Comprehensive integration**: Unlike prior work focusing on isolated components (extraction OR classification OR querying), this research delivers an integrated pipeline from text to answerable knowledge.
2. **Operational focus**: While much research remains academic, this work is explicitly designed for operational deployment in AU-CEWS early warning systems, shaping requirements and design decisions.
3. **Hierarchical classification**: Explicit development of a hierarchical violence taxonomy and corresponding multi-level classification approach enables flexible querying at appropriate granularity.
4. **Knowledge base architecture**: Systematic design of knowledge representation and storage to support not just retrieval but relationship reasoning and pattern analysis.
5. **Natural language accessibility**: Question-answering interface makes extracted intelligence accessible to diverse stakeholders without technical database expertise.
6. **African context adaptation**: Explicit focus on African conflicts, actors, and reporting patterns rather than applying generic systems designed for other contexts.

By addressing the complete pipeline with emphasis on operational utility, this research advances beyond extraction accuracy as the sole metric toward a holistic assessment of how effectively the system supports decision-making for early warning and conflict response.

# 7. Scope and Limitations

This research project has defined scope and acknowledged limitations that bound its objectives and guide realistic expectations for deliverables.

## Scope

### In Scope:

1. **Event Extraction from English Text**: The system will process English-language news articles to extract violent event information. English remains the dominant language for international news aggregation in AU-CEWS systems.
2. **Pre-Collected News Articles**: The research utilizes news articles already aggregated by AU-CEWS Africa Media Monitor. The system will process articles that have been collected rather than implementing news gathering mechanisms.
3. **Violent Events**: The focus is specifically on violent events (attacks, conflicts, repression, terrorism, etc.) rather than all event types. This domain focus enables deeper specialization and higher quality extraction.
4. **5W1H Event Attributes**: The extraction targets Who (perpetrator), What (event type), Whom (victim), Where (location), When (time), and How (method/circumstances) as the core event attributes.
5. **African Geographic Focus**: While the system can process violent events globally, the taxonomy design, training data, and evaluation emphasize African conflicts and violence patterns.
6. **Hierarchical Violence Taxonomy**: Development of a comprehensive, multi-level taxonomy of violent events with 3-5 hierarchical levels and approximately 50-100 distinct event categories at the finest granularity.
7. **Knowledge Base Implementation**: Design and implementation of a structured knowledge base architecture combining relational, graph, and full-text storage for extracted events with their taxonomic classifications and relationships.
8. **Natural Language Question-Answering**: Development of a Q&A interface allowing users to query the knowledge base using natural language questions focused on 5W1H information retrieval.
9. **Prototype System**: Creation of a functional prototype demonstrating the complete pipeline, suitable for evaluation and limited operational testing but not necessarily production-ready at scale.
10. **Offline Processing**: Event extraction and classification occur after articles are collected (batch processing), not in real-time as articles are published.

### Out of Scope:

1. **Real-Time Event Extraction**: The system is not designed for real-time processing with millisecond latency requirements. Processing time per article is expected to be on the order of seconds to minutes.
2. **News Aggregation and Clustering**: Collection of news articles from sources and clustering of articles about the same event are assumed to be handled by existing systems (Africa Media Monitor) and are not part of this research.
3. **Multilingual Processing**: While critically important for comprehensive African monitoring, extracting from Arabic, French, Swahili, Amharic, and other African languages is beyond this research scope. Future extensions should address multilingual capabilities.
4. **Event Prediction and Forecasting**: The system extracts and classifies reported events from news text; it does not predict future events or provide early warning forecasts based on patterns (though the accumulated knowledge base could support such analysis in future work).
5. **Verification and Fact-Checking**: The system treats news articles as ground truth and does not verify accuracy, detect misinformation, or assess source credibility. Manual verification by domain experts remains necessary for operational use.
6. **Sentiment and Impact Analysis**: While extracting what happened, the system does not analyze media framing, public sentiment, or assess humanitarian impact beyond what is explicitly mentioned in text.
7. **Image and Video Analysis**: The system processes only textual content; analysis of images, videos, or other multimedia content in news reports is excluded.
8. **Complex Causal Reasoning**: While the system may extract some explicit causal relationships mentioned in text, complex inference about root causes, motivations, or conflict dynamics is beyond scope.
9. **Production-Scale Deployment**: The deliverable is a functional prototype suitable for evaluation and pilot use; production-level concerns like failover, multi-region replication, enterprise security, and 24/7 operations are not addressed.
10. **Integration with Existing AU Systems**: While APIs will be designed for potential integration, actual deployment into existing AU-CEWS technical infrastructure is beyond this research scope.

## Limitations

### Data Limitations:

1. **Training Data Size**: Despite efforts to compile substantial training data (target: 2,500-3,500 annotated events), this remains modest compared to some NLP tasks. Some rare event types may have insufficient training examples for optimal classification accuracy.
2. **Annotation Consistency**: Inter-annotator agreement, while monitored, will not be perfect. Some subjectivity in human annotation may introduce noise in training data.
3. **News Reporting Bias**: The system learns from news reports, which have inherent biases in geographic coverage (urban areas over-represented), event types (dramatic attacks more likely reported), and perspectives (government sources may be privileged). These biases propagate to system outputs.
4. **Temporal Coverage**: Training data will span 2-3 years; rapidly evolving conflict patterns or entirely new violence forms may not be well-represented.

### Technical Limitations:

1. **Extraction Accuracy**: State-of-the-art event extraction is far from perfect. Expected performance is 70-85% F1-score for complete event extraction, meaning 15-30% of events will be missed or incompletely extracted.
2. **Classification Accuracy**: Hierarchical classification is challenging; expected accuracy decreases at finer-grained levels. Top-level category accuracy may reach 85-90%, but specific event subtypes may achieve only 70-75% accuracy.
3. **Cross-Sentence Information Integration**: Events described across multiple paragraphs or requiring extensive context remain challenging to extract completely.
4. **Rare Event Types**: Events with few training examples (e.g., highly specific attack methods or rare actor types) will have lower classification accuracy.
5. **Ambiguous Cases**: Some events genuinely span multiple categories (e.g., election violence that is simultaneously political violence and communal violence); the system may struggle with these edge cases.

### Linguistic Limitations:

1. **English Only**: As noted, non-English text cannot be processed, severely limiting coverage in Francophone, Lusophone, and Arabophone Africa.
2. **Informal Language**: While focused on news articles with relatively formal language, social media text or very informal reporting styles may challenge the NLP pipeline.
3. **Implied Information**: Events where critical details are implied rather than stated explicitly (e.g., actors identified only by pronoun reference to earlier text) may be incompletely extracted.
4. **Figurative Language**: Metaphors, euphemisms, and indirect language may be misinterpreted (e.g., “the situation escalated” when referring to violence).

### Taxonomic Limitations:

1. **African Context Specificity**: The taxonomy is optimized for African conflicts; application to other regions may require adaptation as some categories may be Africa-specific while categories relevant elsewhere (e.g., certain types of gang violence) may be under-represented.
2. **Taxonomy Evolution**: Conflict and violence continuously evolve; new tactics, technologies, and forms of violence emerge. The taxonomy will require ongoing maintenance and extension.
3. **Boundary Cases**: Some events will be difficult to classify into distinct categories; taxonomic boundaries cannot perfectly capture the fluid reality of violence.
4. **Granularity Tradeoffs**: The taxonomy must balance specificity (enabling detailed analysis) against reliability (more specific categories are harder to classify accurately). The chosen granularity represents a compromise.

### Knowledge Base Limitations:

1. **Relationship Inference**: While some relationships (temporal, spatial) can be automatically inferred, complex relationships (causal links, coordinated campaigns) often require human judgment.
2. **Entity Resolution**: Linking mentions to the same real-world actor or location across events is error-prone, especially with name variations and ambiguous references.
3. **Data Quality Dependency**: Knowledge base quality depends entirely on extraction quality; errors in extraction propagate throughout the knowledge base.
4. **Scalability**: While designed for growth, performance may degrade with millions of events without additional optimization.

### Question-Answering Limitations:

1. **Question Complexity**: The Q&A system handles factual 5W1H questions effectively but struggles with complex analytical questions requiring inference, comparison across many dimensions, or causal reasoning.
2. **Natural Language Understanding**: Ambiguous questions, unconventional phrasings, or questions using terminology not in the system’s vocabulary may be misinterpreted.
3. **Answer Completeness**: Answers aggregate information from the knowledge base but cannot provide context or analysis beyond what was extracted.
4. **No Explanatory Capability**: The system retrieves answers but does not explain why events occurred, their significance, or implications for policy.

### Operational Limitations:

1. **Processing Latency**: Each article requires seconds to minutes to process; bulk processing of large article backlogs takes hours to days.
2. **Human Review Requirement**: For operational use, extracted events should be reviewed by human analysts before use in decision-making, reducing but not eliminating the human workload.
3. **Continuous Maintenance**: The system requires ongoing maintenance: updating training data for new violence patterns, refining the taxonomy, tuning classification models, and fixing errors.
4. **Expertise Required**: While the Q&A interface is user-friendly, system administration, model retraining, and taxonomy updates require technical expertise in NLP and machine learning.

### Ethical Limitations:

1. **Reproduction of Biases**: The system may reproduce and amplify biases present in news reporting, potentially over-representing violence involving certain actors or regions while underrepresenting others.
2. **De-contextualization Risk**: Structured event records may strip away important context, nuance, and human impact, potentially facilitating technocratic responses that miss deeper issues.
3. **Dual-Use Concerns**: Event extraction technologies could potentially be misused for surveillance, targeting, or censorship; appropriate governance frameworks must accompany deployment.

Despite these limitations, the research makes significant contributions to automated violent event intelligence for early warning systems. By explicitly acknowledging limitations, the research establishes realistic expectations and identifies directions for future improvement. The integrated approach—combining extraction, hierarchical classification, knowledge base organization, and natural language querying—represents a substantial advancement in making violent event information accessible and actionable for decision-makers, even with these acknowledged constraints.

# 8. Application of Results

The outputs of this research—an operational event extraction system, hierarchical violence taxonomy, structured knowledge base, and natural language query interface—have significant practical applications across multiple stakeholder communities and use contexts. The value extends from immediate operational utility to long-term strategic analysis and policy development.

## Operational Early Warning and Crisis Response

### African Union Continental Early Warning System (AU-CEWS)

As the primary stakeholder, AU-CEWS gains multiple operational capabilities:

1. **Accelerated Incident Awareness**: Rather than manually reading hundreds of daily news articles, analysts can query “What violent events occurred in the Sahel region in the past 24 hours?” and immediately see structured summaries, enabling faster situation awareness.
2. **Systematic Event Documentation**: Every extracted event is documented with standardized 5W1H attributes and taxonomic classification, creating consistent records that facilitate cross-event comparison and longitudinal analysis.
3. **Geographic Hotspot Identification**: Queries like “Which districts in Ethiopia had more than 3 violent events last month?” quickly identify geographic concentrations requiring attention, supporting resource allocation and deployment decisions.
4. **Actor Tracking**: The knowledge base enables tracking specific armed groups, militias, or terrorist organizations across time and space: “Show me all attacks attributed to Al-Shabaab in the past quarter” reveals operational patterns, target preferences, and geographic reach.
5. **Event Type Trend Analysis**: Hierarchical classification enables questions at multiple granularities: “Is terrorism increasing in West Africa?” (broad category) or “Are suicide bombings specifically becoming more common?” (specific tactic). This supports threat assessment and prioritization.
6. **Rapid Reporting Generation**: When senior leadership requests a briefing on recent violence in a region, analysts can query the knowledge base and generate structured reports in minutes rather than hours or days.
7. **Early Warning Indicators**: Accumulation of events over time enables identification of escalation patterns: sudden spikes in violence, emergence of new actors, shifts in tactics, or geographic expansion of conflicts all become visible in aggregate query results.

### Regional Economic Communities (RECs) Early Warning Systems

AU member Regional Economic Communities (ECOWAS, IGAD, SADC, etc.) operating their own early warning mechanisms benefit from similar capabilities adapted to their regional focus. The system can filter to specific regions (“Show violent events in ECOWAS member states”) and support regional reporting requirements.

### National Early Warning Centers

Individual African countries with national early warning capacities can deploy adapted versions focusing on domestic monitoring. For example, Kenya’s National Counter-Terrorism Center could track terrorism-specific events, while Ethiopia’s early warning system might emphasize communal violence categories.

## Humanitarian Response and Protection

### Humanitarian Organizations

International NGOs, UN agencies (OCHA, UNHCR, WFP), and the Red Cross/Red Crescent movement require detailed information about violence affecting civilian populations:

1. **Protection Planning**: Queries like “What violence against civilians occurred in displacement camp regions?” or “Where have aid workers been kidnapped in the past month?” directly inform protection strategies and safety protocols.
2. **Needs Assessment**: Understanding what types of violence affected which populations (attacks on villages, forced displacement, sexual violence) helps estimate humanitarian needs: shelter, medical care, psychosocial support, livelihood assistance.
3. **Access and Security**: The knowledge base enables mapping of violence along humanitarian access routes, identifying checkpoints where aid convoys face risks, and tracking attacks on humanitarian assets.
4. **Advocacy**: Systematic documentation of violence patterns provides evidence for advocacy with governments, armed groups, and international bodies regarding protection of civilians and humanitarian access.

### International Committee of the Red Cross (ICRC)

ICRC’s mandate to monitor violations of international humanitarian law (IHL) could be supported through queries identifying specific violation types: attacks on medical facilities, targeting of civilians, use of prohibited weapons, etc.

## Security and Defense Analysis

### African Union Peace Support Operations

AU peacekeeping missions (AMISOM, MINUSMA successors, etc.) benefit from understanding violence in their areas of operation:

1. **Threat Assessment**: Before deploying to new areas, missions can query historical violence patterns to assess risks and plan force protection measures.
2. **Operational Planning**: Identifying where different types of violence concentrate (ambushes along roads, attacks on government facilities, IED threats) informs patrol routes, convoy procedures, and base security.
3. **Mission Reporting**: Peacekeeping missions must report on security situations; the knowledge base provides systematic context and comparison to previous periods.

### National Defense and Intelligence Services

African military and intelligence services can use the system for open-source intelligence (OSINT) to understand non-state armed groups, track cross-border threats, and monitor conflict dynamics in neighboring countries.

### International Partners

Organizations like NATO, UN DPPA (Department of Political and Peacebuilding Affairs), and national intelligence services supporting African security cooperation can use extracted information to understand conflict patterns and inform assistance strategies.

## Conflict Research and Analysis

### Academic Research Community

Scholars studying African conflicts gain access to structured, queryable data for research:

1. **Quantitative Conflict Studies**: The knowledge base provides datasets for statistical analysis of violence patterns, testing theories about conflict causes and dynamics, and evaluating intervention effectiveness.
2. **Comparative Analysis**: Researchers can compare violence patterns across countries, regions, or time periods: “How do election-related violence patterns differ between East and West Africa?”
3. **Actor Analysis**: Studies of non-state armed groups, state repression, or communal violence benefit from comprehensive actor-coded event data.
4. **Methodological Advancement**: The annotated dataset and extraction system contribute resources for NLP and computational social science research.

### Think Tanks and Policy Institutes

Organizations like the Institute for Security Studies (ISS), International Crisis Group, or Africa Center for Strategic Studies produce policy analysis and recommendations informed by systematic violence data from the knowledge base.

### Conflict Early Warning and Forecasting Research

While this system doesn’t predict future events, accumulated historical data in the knowledge base provides the foundation for machine learning models that do: identifying precursor patterns, estimating escalation probabilities, and forecasting geographic expansion of conflicts.

## Governance and Rule of Law

### Transitional Justice Mechanisms

Truth commissions, special courts, and accountability mechanisms investigating past violence can use the knowledge base to systematically document atrocities, identify patterns of violations, and locate witnesses.

### International Criminal Court (ICC) and Hybrid Courts

Investigations of war crimes, crimes against humanity, and genocide require systematic evidence collection. The knowledge base provides initial documentation of alleged crimes, though legal prosecutions require verification beyond news reports.

### National Justice Systems

Domestic prosecutions of terrorism, organized crime, or human rights violations can draw on systematically documented events, timelines, and actor networks from the knowledge base.

## Development and Peacebuilding

### Development Organizations

World Bank, African Development Bank, and bilateral development agencies integrating conflict sensitivity into development programming benefit from understanding violence patterns in project areas:

1. **Risk Assessment**: Before investing in infrastructure or social programs, queries reveal historical violence that might threaten projects.
2. **Conflict-Sensitive Design**: Understanding local conflict dynamics (communal tensions, resource conflicts) informs project design to avoid exacerbating divisions.
3. **Evaluation**: Post-intervention evaluation can assess whether development programs in conflict-affected areas correlated with violence reduction.

### Peacebuilding Organizations

Organizations implementing mediation, dialogue, reconciliation, or community peace initiatives use the knowledge base to:

1. **Context Analysis**: Understanding historical violence patterns between communities or regions informs intervention design.
2. **Baseline Establishment**: Documenting violence before interventions enables impact evaluation afterward.
3. **Early Warning for Peacebuilders**: Monitoring recent events helps peacebuilding staff anticipate tensions and adjust programs.

## Media and Public Information

### Journalists and Media Organizations

Journalists covering African conflicts can query the knowledge base for context, historical patterns, and comparative statistics to enrich reporting with data-driven insights.

### Fact-Checking Organizations

Organizations verifying claims about violence can cross-reference allegations against the knowledge base’s systematically extracted information.

### Public Communication and Civic Education

Civil society organizations educating publics about conflict patterns, peace processes, or transitional justice can use aggregated statistics and timelines from the knowledge base.

## Advanced Analytical Applications

### Network Analysis

The knowledge base’s relationship structure enables sophisticated network analysis:

1. **Actor Networks**: Identifying which armed groups cooperate, which are rivals, and how alliances evolve.
2. **Retaliatory Cycles**: Tracing chains of attacks and reprisals to understand conflict escalation dynamics.
3. **Geographic Diffusion**: Modeling how violence spreads spatially through regions.

### Temporal Pattern Mining

Queries identifying periodic patterns (violence spikes around elections, religious holidays, or harvests), trend analysis (which violence types increasing/decreasing), and seasonality.

### Predictive Analytics Foundation

While not directly providing predictions, the accumulating historical knowledge base enables training of:

1. **Conflict Forecasting Models**: Predicting geographic or temporal patterns of future violence.
2. **Early Warning Indicators**: Identifying precursor patterns that signal escalation.
3. **Scenario Analysis**: Estimating how different policy choices might affect violence trajectories.

## System Scalability and Adaptation

### Adaptation to Other Event Types

The methodology developed for violent events can be adapted to extract other event types relevant to early warning and development:

1. **Political Events**: Elections, protests, government changes, policy decisions.
2. **Humanitarian Crises**: Floods, droughts, disease outbreaks, famines.
3. **Economic Events**: Market disruptions, infrastructure failures, strikes.
4. **Environmental Events**: Resource conflicts, climate change impacts, environmental degradation.

### Geographic Extension

While designed for Africa, the methodology can be adapted to other regions by:

1. **Taxonomy Revision**: Adjusting event categories for different conflict patterns.
2. **Training Data Collection**: Annotating news from new regions.
3. **Entity Recognition**: Adapting to new actor names and place names.

### Temporal Extension

The knowledge base accumulates value over time:

1. **Historical Baselines**: Years of data establish normal patterns, making anomalies detectable.
2. **Longitudinal Studies**: Long-term trends become analyzable only with consistent multi-year data.
3. **Conflict Life Cycle Analysis**: Observing conflicts from emergence through escalation to resolution (or transformation).

## Operational Impact Assessment

### Quantifiable Benefits

1. **Time Savings**: Analysts estimate manually processing daily news for violent events requires 2-3 hours daily. Automated extraction with query interface reduces this to 30 minutes of targeted querying and verification—approximately 75% time savings.
2. **Coverage Expansion**: Manual analysis typically covers only major events from prominent sources. Automated processing can handle 10x more articles, capturing lower-profile events in peripheral regions.
3. **Response Speed**: Accelerating awareness from 24-48 hours (manual review) to 2-3 hours (automated extraction) improves crisis response timeliness.
4. **Consistency**: Standardized extraction and classification eliminates analyst-to-analyst variability in event coding.

### Qualitative Benefits

1. **Cognitive Load Reduction**: Freeing analysts from tedious reading and coding enables focus on interpretation, strategy, and decision support.
2. **Data-Driven Insights**: Systematic data enables statistical rigor previously impossible with manual methods.
3. **Institutional Memory**: Knowledge base preserves historical information beyond individual analysts’ recall, surviving staff turnover.
4. **Democratization of Access**: Natural language querying enables non-technical stakeholders (policy officials, field staff) to directly access information without analyst intermediation for every question.

## Long-Term Strategic Value

### Policy Evidence Base

Accumulating years of systematically documented violence creates an evidence base for:

1. **Intervention Evaluation**: Did peacekeeping deployments reduce violence? Did peace agreements hold?
2. **Comparative Policy Studies**: Which approaches to counterterrorism, conflict resolution, or violence prevention correlate with reduced violence?
3. **Resource Allocation**: Evidence-based prioritization of early warning, peacekeeping, or development resources toward highest-risk regions.

### Academic Contribution

The annotated dataset, taxonomy, and extracted knowledge base constitute valuable resources for:

1. **Conflict Research**: Enabling quantitative studies previously impossible due to data limitations.
2. **NLP Advancement**: Contributing to event extraction, domain-specific NLP, and African language/context NLP research.
3. **Methodology Sharing**: Publishing methods enables replication and adaptation by other researchers and practitioners.

### Capacity Building

The research process—particularly taxonomy development and data annotation—builds capacity among African researchers, AU staff, and early warning professionals in:

1. **Conflict Analysis**: Systematic frameworks for understanding and categorizing violence.
2. **Data Science**: Experience with NLP, machine learning, and knowledge engineering.
3. **Research Methodology**: Rigorous approaches to studying conflict.

## Ethical Considerations and Responsible Use

While highlighting applications, it’s essential to acknowledge ethical responsibilities:

1. **Do No Harm**: Information about violence must not facilitate targeting of individuals or communities.
2. **Privacy Protection**: While events are public (from news), care must be taken not to expose survivors or witnesses to additional risk.
3. **Bias Awareness**: Users must understand that the system reflects news media biases and doesn’t provide complete or unbiased violence accounting.
4. **Human Judgment**: Automated extraction assists but doesn’t replace human analysis; critical decisions must incorporate human expertise and contextual understanding.
5. **Transparency**: System capabilities and limitations should be clearly communicated to prevent misuse or over-reliance.

By serving these diverse stakeholders and applications, the research delivers practical value extending far beyond academic contribution, supporting better-informed decision-making for peace, security, and protection of civilians across Africa.

# 9. Annexes

## Project Timeline

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Time (in weeks) | | | | | | | | | | | | | | | | | |
| No | Activities | Mar | | | | Apr | | | | May | | | | Jun | | | |
| w1 | w2 | w3 | w4 | w1 | w2 | w3 | w4 | w1 | w2 | w3 | w4 | w1 | w2 | w3 | w4 |
| 1 | Literature Review |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2 | Methodology |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3 | Development of Software |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4 | Testing of developed system |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5 | Thesis writing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 6 | Incorporate comments from advisors |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 7 | Final Submission |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# 10. References

[1] Tanev, H., Atkinson, M. & Piskorski, J., 2008. Real-Time News Event Extraction for Global Crisis Monitoring. *13th International Conference on Natural Language and Information Systems: Application of Natural Language to Information Systems (NLDB 2008). Lecture Notes in Computer Science, vol 5039,* pp. 207-218.

[2] Hogenboom, F., Frasincar, F., Kaymak, U. & De Jong, F., 2011. An overview of event extraction from text. *In Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011),* Volume 779, pp. 48-57.

[3] Ikonomakis, M., Kotsiantis, S. & Tampakas, V., 2005. Text Classification Using Machine Learning Techniques. *WSEAS Transactions on Computers,* 4(8), pp. 966-974.

[4] Manning, C. D. & Schutze, H., 1999. *Foundations of Statistical Natural Language Processing.* 1st ed. s.l.:MIT Press.

[5] Hogenboom, F. et al., 2016. A survey of event extraction methods from text for decision support systems. *Decision Support Systems,* Volume 85, pp. 12-22.

[6] CEWS, 2013. *Data Collection and Analysis Tools for Continental Early Warning System.* Addis Ababa: Unpublished Internal Document.

[7] Offermann, P., Levina, O., Schönherr, M. & Bub, U., 2009. Outline of a design science research process. *In Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology,* Volume ACM, pp. 7-18.

[8] Taye Abdulkadir Edris & Roopesh Kevin Sungkur (2019) Knowledge Discovery from Free Text: Extraction of Violent Events in the African Context, *New Review of Information Networking,* 24:2, 153-177, DOI: 10.1080/13614576.2019.1615529

[9] Farzindar, A. and Khreich, W., 2015. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence,* 31(1), pp. 132-164.

[10] Suchanek, F. M., Kasneci, G. and Weikum, G., 2007. Yago: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. *In Proceedings of the 16th International Conference on World Wide Web,* pp. 697-706, ACM.

[11] Hienert, D. and Luciano, F., 2015. Extraction of Historical Events from Wikipedia. *ESWC 2012 Satellite Events, LNCS 7540,* pp. 16-28.

[12] Magnuson, A., Dialan, V. and Mallela, D., 2015. Event Recommendation using Twitter Activity. *RecSys’15,* ACM, 978-1-4503-3692-5/15/09, September 16-20, 2015, Vienna, Austria.

[13] Becker, H., Iter, D., Naaman, M. and Gravano, L., 2012. Identifying Content for Planned Events Across Social Media Sites. *WSDM’12,* ACM 978-1-4503-0747-5/12/02, Seattle, Washington, USA.

[14] Piskorski, J., Tanev, H., and Wennerberg, P. O., 2007. Extracting Violent Events from On-Line News for Ontology Population. *BIS 2007, LNCS 4439,* pp. 287-300.

[15] Wang, W. and Zhao, D., 2012. Chinese News Event 5W1H Semantic Elements Extraction for Event Ontology Population. *International World Wide Web Conference Committee (IW3C2), WWW 2012 Companion,* April 16-20, 2012, Lyon, France.

[16] Raleigh, C., Linke, A., Hegre, H. and Karlsen, J., 2010. Introducing ACLED: An Armed Conflict Location and Event Dataset. *Journal of Peace Research,* 47(5), pp. 651-660.

[17] Sundberg, R. and Melander, E., 2013. Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research,* 50(4), pp. 523-532.

[18] Leetaru, K. and Schrodt, P. A., 2013. GDELT: Global Data on Events, Location, and Tone, 1979-2012. *ISA Annual Convention,* Volume 2, Number 4, pp. 1-49.